

# An Approach for Weakly-Supervised Deep Information Retrieval

Sean MacAvaney, Kai Hui, Andrew  
Yates

July 15, 2018

(Appeared at neu-IR workshop at SIGIR '17)



Recent improvements to IR using neural approaches [1-4]

But neural approaches benefit from having lots of data

Recent work [5] generated pseudo-labels using a query log and corpus

[1] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Cro . 2016. A deep relevance matching model for ad-hoc retrieval. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 55–64.

[2] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. A Position-Aware Deep Model for Relevance Matching in Information Retrieval. arXiv preprint arXiv:1704.03940 (2017).

[3] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to Match Using Local and Distributed Representations of Text for Web Search. In Proceedings of WWW 2017. ACM.

[4] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2016. A Study of MatchPyramid Models on Ad-hoc Retrieval. CoRR abs/1606.04648 (2016).

[5] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Cro . 2017. Neural Ranking Models with Weak Supervision (SIGIR '17).

# Motivation



We present an approach that eliminates the need for a query log.

We present an approach that eliminates the need for a query log.

**Main idea:** Use news articles, with headlines acting as pseudo-queries, and the article body acting as pseudo-relevant documents.

This can work because headlines are often short descriptions of the article.

## May Wedding Set For Laura Myers

Published: January 11, 1987

The engagement of Laura Susan Myers to Steven Hammond Poppe, the son of Mr. and Mrs. Fred C. Poppe of West Islip, L.I., has been announced by Mr. and Mrs. W. Earle Myers of Long Valley, N.J., parents of the bride-to-be. A May wedding is planned.

Ms. Myers, an account executive at the Direct Marketing Association in New York, graduated from Hollins College. She is a granddaughter of the late Ralph Alexander Dickson of Gastonia, N.C., the founder and owner of Genuine Parts Inc. of North Carolina a principal in the electronics firm of Shiflett & Dickson. and of the late O. W. Myers.

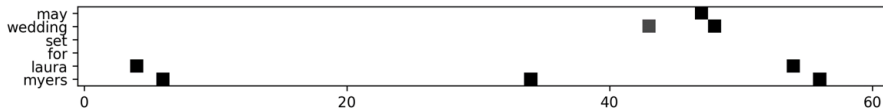
# Example

## May Wedding Set For Laura Myers

Published: January 11, 1987

The engagement of Laura Susan Myers to Steven Hammond Poppe, the son of Mr. and Mrs. Fred C. Poppe of West Islip, L.I., has been announced by Mr. and Mrs. W. Earle Myers of Long Valley, N.J., parents of the bride-to-be. A May wedding is planned.

Ms. Myers, an account executive at the Direct Marketing Association in New York, graduated from Hollins College. She is a granddaughter of the late Ralph Alexander Dickson of Gastonia, N.C., the founder and owner of Genuine Parts Inc. of North Carolina a principal in the electronics firm of Shiflett & Dickson. and of the late O. W. Myers.



# Challenges



Hard-Negative Problem

Mismatched Interaction Problem

# Hard-Negative Problem



How to choose *negative* training samples?



# Hard-Negative Problem



How to choose *negative* training samples?

We query articles from the corpus using the headline and pick the top  $n$  results as negative training samples. (BM25)

# Mismatched Interaction Problem



## Headlines do not always act as good queries (e.g. figurative language)

### SPORTS OF THE TIMES; When Bird Flies In

By Ira Berkow  
Published: November 12, 1987

LARRY BIRD was in town the other day, and left, but chances are he'll be heard from again.

He didn't arrive in disguise, and he didn't leave in disguise. Not quite. His game's the same - fill the hoop, fill the lane, fill the other team with vexation. The trappings, however, are a little less familiar.

He's sleeker, though not sleek; he's thinner, though not thin; he's more muscular, but not quite muscular.

There's a reason for this physical change in the man many have called the best all-round player to ever heave a ball at a hoop.

when bird flies in

All Videos Shopping Images News More

About 3.270.000 results (0,74 seconds)

#### Birds as Omens and Signs | Exemple.com

<https://exemple.com> › Animal Guides ▾

May 25, 2017 - When a bird flies into your car or house window and is knocked out or die from the impact. This may be an omen that you ...

#### Explore Bird Superstitions and Myths | Exemple.com

<https://exemple.com> › Animal Guides ▾

Nov 4, 2015 - A lot of these superstitions involve luck, both good or bad luck. For example, a bird flies into your house signifies that an important message ...

#### Ornithomancy: Divination from the Flight and Cries of Birds | Exemple.com

<https://exemple.com> › Fortune Telling & Divination ▾

Apr 18, 2016 - Birds are majestic creatures and have long fascinated humankind as spirit world. With the ability to walk on land and fly

## Approach 1: Ranking filter

Discard article if not retrieved in top BM25 results

This eliminates articles like *When Bird Flies In.*

## **Approach 2: Interaction filter**

Only pick training samples that have similar document interactions to a set of template interactions

Since models “see” the data differently, interaction filters are model-specific

## General interaction filter

- 1 Use a set of “template” query-document pairs
- 2 For each template, calculate a *mock interaction embedding* (model-specific function  $m(p)$ )
- 3 Keep only the top  $n$  most similar articles (given model-specific distance function  $d(m_1, m_2)$ )

## Example interaction filter (trivial)

$$m(p) = \frac{\# \text{ matching query terms in document}}{\# \text{ query terms}}$$

$$d(m_1, m_2) = |m_1 - m_2|$$

## PACRR interaction filter

$m$ : maximum value for each query term in query-document similarity matrix

$d$ : “aligned” mean squared error, i.e. minimum MSE over each possible alignment of mock embeddings

### Kept

nbc to buy miami tv outlet  
state looks to congress again for highway aid  
151 new yorkers get science honor

### Discarded

when allies don't see eye to eye  
wars can't be won only from above  
diseases common in ashkenazim may be random

# Experiment

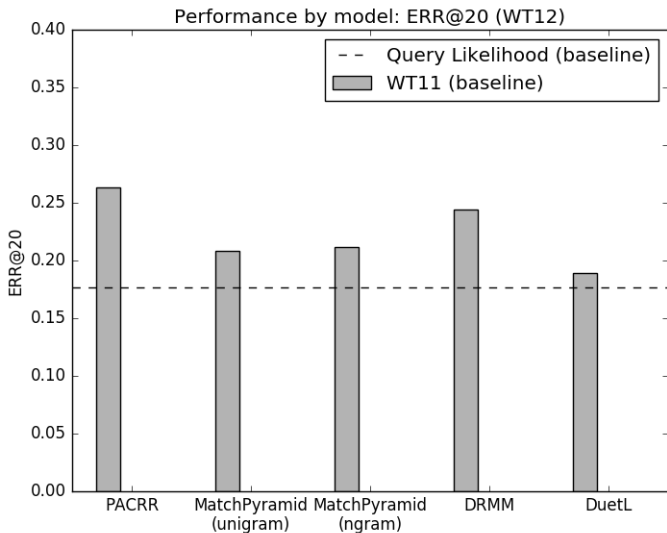
Train 5 neural IR models using New York Times Corpus data

Evaluate on TREC Web 2012-2014 with ERR@20 and nDCG@20

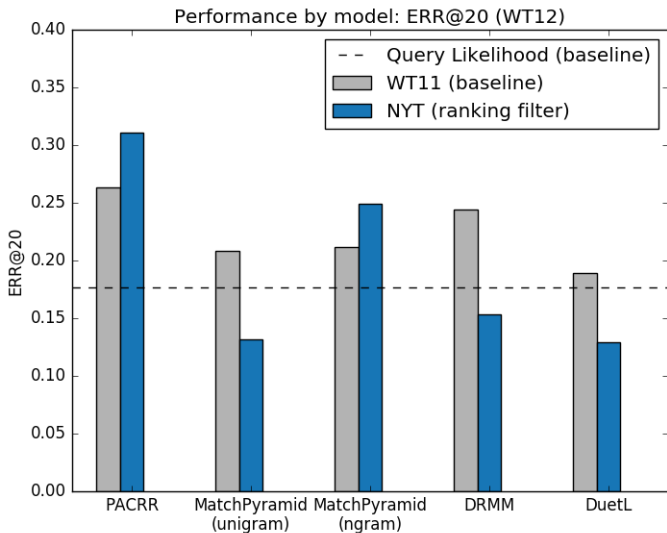
Baselines: Query Likelihood; each model trained on TREC Web 2011 data



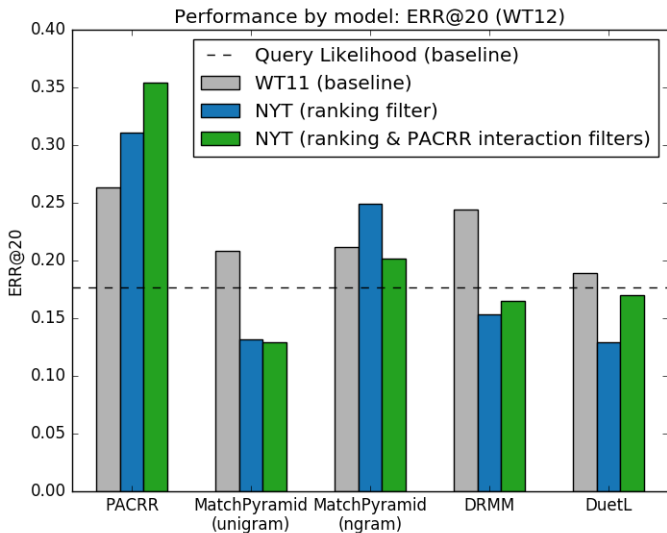
# Results



# Results



# Results



# Conclusions

We showed promising results for using news articles for neural IR training

## **Future Work:**

Interaction filters for other models

Use a different source of interaction templates

Questions?

# Mismatched Interaction Problem

## PACRR

